Analysing grouping of nucleotides in DNA sequences using lumped processes constructed from Markov chains

Yann Guédon, Y. d'Auberton-Carafa, C. Thermes

Analysing grouping of nucleotides in DNA sequences using lumped processes constructed from Markov chains

Yann Guédon[1*], Yves d'Aubenton[2] and Claude Thermes[2]

[1] *Unité Mixte de Recherche CIRAD/CNRS/INRA/IRD/Université Montpellier II*
Botanique et Bioinformatique de l'Architecture des Plantes,
TA 40/PS2, 34398 Montpellier Cedex 5, France
E-mail: guedon@cirad.fr
Phone: (33) 4 67 61 65 78 Fax : (33) 4 67 61 56 68

[2] *Centre de Génétique Moléculaire, CNRS*
91198 Gif sur Yvette Cedex, France
E-mail: daubenton@cgm.cnrs-gif.fr, thermes@cgm.cnrs-gif.fr

Running head: Lumped processes for analysing grouping in DNA sequences

Analysing grouping of nucleotides in DNA sequences using lumped processes constructed from Markov chains

Keywords: DNA sequence; Lumpability; Markov chain; Model selection.

Abstract

The most commonly used models for analysing local dependencies in DNA sequences are (high-order) Markov chains. Incorporating knowledge relative to the possible grouping of the nucleotides enables to define dedicated sub-classes of Markov chains. The problem of formulating lumpability hypotheses for a Markov chain is therefore addressed. In the classical approach to lumpability, this problem can be formulated as the determination of an appropriate state space (smaller than the original state space) such that the lumped chain defined on this state space retains the Markov property. We propose a different perspective on lumpability where the state space is fixed and the partitioning of this state space is represented by a one-to-many probabilistic function within a two-level stochastic process. Three nested classes of lumped processes can be defined in this way as sub-classes of first-order Markov chains. These lumped processes enable parsimonious reparameterizations of Markov chains that help to reveal relevant partitions of the state space. Characterizations of the lumped processes on the original transition probability matrix are derived. Different model selection methods relying either on hypothesis testing or on penalized log-likelihood criteria are presented as well as extensions to lumped processes constructed from high-order Markov chains. The relevance of the proposed approach to lumpability is illustrated by the analysis of DNA sequences. In particular, the use of lumped processes enables to highlight differences between intronic sequences and gene untranslated region sequences.

1 Introduction

Statistical and probabilistic properties of patterns (also called words or motifs) have been of considerable interest in DNA sequence analysis; see Reinert *et al.* (2000) for a review. Most of the proposed methods rely on (high-order) Markov chains that represent the local dependencies in the sequences. In this framework, various methods have been proposed to find patterns with unexpected frequencies (Prum *et al.*, 1995) or to study the distances between two successive occurrences of a given pattern (Robin and Daudin, 1999; Stefanov, 2003). It is also well know that the class of high-order Markov chains is not structurally rich (Bühlmann and Wyner, 1999), which implies that any kind of parsimonious parameterizations of local dependencies is not possible. This point is illustrated in Table 1 which shows the number of free parameters of four-state Markov chains as a function of their order. The very "discontinuous" increase of the number of free parameters causes the estimated high-order Markov chains to be generally overparameterized. This drawback can be overcome by incorporating structure in order to define sub-classes of high-order Markov chains. This point is notably illustrated by variable-order Markov chains (Weinberger *et al.*, 1995; Ron *et al.*, 1996; Bühlmann and Wyner, 1999) where the order, or

equivalently the memory length, is variable (and depends on the context) instead of being fixed; see Mächler and Bülhmann (2004) for an example of DNA sequence analysis using variable-order Markov chains. We will here investigate another approach which consists in taking into account some knowledge relative to state grouping. In the case of DNA sequences (four-state sequences where the states are the nucleotides A, C, G and T), grouping may be related either to the purine-pyrimidine (A/G, C/T), to the strong-weak hydrogen bonding (G/C, A/T) or to the keto-amino (T/G, A/C) classifications. A possible grouping can be illustrated by the human intron example. The estimated transition probability matrix for the original four-state space (Table 2) shows that the two rows corresponding to the purines A and G are close. This suggests that the four-state first-order Markov chain is overparameterized. By grouping A and G into a single state A/G that denotes purines, a lumped process can be estimated which achieves a better tradeoff between the fit to the data and the parsimony of the model as shown in Section 6.1.

In the classical approach to lumpability (Burke and Rosenblatt, 1958; Kemeny and Snell, 1976), the problem of formulating lumpability hypotheses for a Markov chain can be stated as the determination of an appropriate state space (smaller than the original state space) such that the lumped chain defined on this state space retains the Markov property. It is known that lumpability can be characterized on the transition probability matrix since the sums of transition probabilities from each state in a partition group to all states of a fixed partition group must be equal (row sum criterion). The applications targeted in this paper should be clearly contrasted with the determination of an appropriate partition of a large state space that may arise in performance and reliability modelling, manpower planning and other practical settings; see Rogers and Plante (1993) whose first part is a tutorial. Our focus is on small but 'dense' (in the sense of having no or a few zero entries) transition probability matrices, not on sparse but large transition probability matrices. Hence, the objective of reducing the state space becomes less relevant and our objective is rather to find a parsimonious parameterization of local dependencies in discrete sequences that help to reveal relevant underlying partitions of the state space. We are thus naturally faced with a model selection problem.

In the proposed approach to lumpability, the state space is fixed and the partitioning of this state space is represented by a one-to-many probabilistic function within a two-level stochastic process. A two-level stochastic process is a pair of stochastic processes where the 'output' process is related to the 'state' process, which is a Markov chain, by a probabilistic function. The most well-known family of two-level stochastic processes are hidden Markov chains (Ephraim and Mehrav, 2002) where the state-to-output mapping is such that a given output may be observed in different states. In the following, we will restrict our attention to two-level stochastic processes with discrete-valued output processes where the state-to-output mapping is such that a given output is observed in a single state and investigate how this type of process can be utilized to formulate lumpability hypotheses. Lumped processes are closely related to variable-order Markov chains. For instance, two of the lumped processes proposed in this paper can be interpreted as a mixed zero-order first-order Markov chain (see Section 3). Both lumped processes and variable-order Markov chains define structurally rich classes of statistical models

with reference to (high-order) Markov chains on the basis of a lumping mechanism that applies to state in the case of lumped processes, and to memories in the case of variable-order Markov chains.

The remainder of this paper is organized as follows. The basis of the classical approach to lumpability is presented in Section 2 with particular emphasis on its translation in terms of two-level stochastic processes. The three possible classes of lumped processes based on first-order Markov chains are derived in Section 3 and the adaptation to our context of the model selection methods used for determining the order of a Markov chain is presented in Section 4. The derivation of lumped processes based on high-order Markov chains is outlined in Section 5. The proposed approach to lumpability is applied to the analysis of DNA sequences in Section 6. Section 7 consists of concluding remarks.

2 Classical lumpability property

In the following, we consider a time-homogeneous first-order Markov chain $\{S_t; t = 0, 1, \ldots\}$ with finite state space $A = \{0, \ldots, N-1\}$ and we assume that all the transitions are possible ($p_{mn} > 0;\ m, n = 0, \ldots, N-1$). The Markov chain $\{S_t\}$ is strongly lumpable with respect to a partition $\widetilde{A} = \left\{\widetilde{A}(0),\ \ldots,\ \widetilde{A}(J-1)\right\}$ of A ($J < N$) if, for every initial state distribution π, the resulting chain $\left\{\widetilde{S}_t\right\}$ is Markovian and the transition probabilities $\widetilde{p}_{ij}$ are invariant under choices of π. In what follows, since we consider only strong lumpability (and not 'weak' lumpability (Kemeny and Snell, 1976)), we shall omit the qualifier. A necessary and sufficient condition for $\{S_t\}$ to be lumpable with respect to a partition $\widetilde{A}$ of A is that, for each pair $\left(\widetilde{A}(i), \widetilde{A}(j)\right)$, the probability of transition from m to some $n \in \widetilde{A}(j)$ is the same for each $m \in \widetilde{A}(i)$ (Kemeny and Snell, 1976, theorem 6.3.2). If this condition is satisfied, the lumped chain $\left\{\widetilde{S}_t\right\}$ is Markovian with transition probabilities $\widetilde{p}_{ij}$ where for each $m \in \widetilde{A}(i)$,

$$\widetilde{p}_{ij} = \sum_{n \in \widetilde{A}(j)} p_{mn}. \qquad \text{(row sum criterion)} \tag{2.1}$$

The classical lumpability property can be translated in terms of two-level stochastic processes as follows (Ephraim and Mehrav, 2002). Let $\{Y_t\}$ be a stochastic process generated by the original Markov chain $\{S_t\}$. Let h be a many-to-one function from A into $\widetilde{A}$ such that $Y_t = h(S_t)$; hence, $Y_t = y_t$ if and only if $s_t \in \widetilde{A}(y_t)$. The deterministic function h may collapse one or more states of the original Markov chain $\{S_t\}$ onto a single output of $\{Y_t\}$. The lumped process $\{S_t, Y_t\}$ can be reparameterized as the lumped Markov chain $\left\{\widetilde{S}_t\right\}$ if the row sum criterion (2.1) is satisfied.

The conditional independence assumptions within the two-level stochastic process $\{S_t, Y_t\}$ translate into the following factorization

$$
\begin{aligned}
& P\left(Y_0 = y_0, \ldots, Y_{\tau-1} = y_{\tau-1}\right) \\
= & \sum_{s_0 \in \widetilde{A}(y_0), \ldots, s_{\tau-1} \in \widetilde{A}(y_{\tau-1})} P\left(Y_0 = y_0, S_0 = s_0, \ldots, Y_{\tau-1} = y_{\tau-1}, S_{\tau-1} = s_{\tau-1}\right) \\
= & \sum_{s_0 \in \widetilde{A}(y_0), \ldots, s_{\tau-1} \in \widetilde{A}(y_{\tau-1})} P\left(S_0 = s_0\right) P\left(S_1 = s_1 | S_0 = s_0\right) \ldots P\left(S_{\tau-1} = s_{\tau-1} | S_{\tau-2} = s_{\tau-2}\right), \quad (2.2)
\end{aligned}
$$

where $\sum_{s_0 \in \widetilde{A}(y_0), \ldots, s_{\tau-1} \in \widetilde{A}(y_{\tau-1})}$ means sum on every possible state sequence of length τ such that $s_0 \in \widetilde{A}\left(y_0\right), \ldots, s_{\tau-1} \in \widetilde{A}\left(y_{\tau-1}\right)$. Since the function h is deterministic, we have for each s_t, $P\left(Y_t = y_t | S_t = s_t\right) = 1$ and the observation probabilities cancel out in the factorization.

If the row sum criterion (2.1) is satisfied, expression (2.2) can be rewritten as a simple product of terms, the first being a sum of initial probabilities and the subsequent terms being sums of transition probabilities

$$
P\left(Y_0 = y_0, \ldots, Y_{\tau-1} = y_{\tau-1}\right) = \left(\sum_{s_0 \in \widetilde{A}(y_0)} \pi_{s_0} \right) \prod_{t=1}^{\tau-1} \sum_{s_t \in \widetilde{A}(y_t)} p_{s_{t-1} s_t}.
$$

3 Lumped processes

We now adopt a different mode of construction of lumped processes. Let $\{X_t\}$ be a stochastic process generated by a stationary Markov chain $\left\{\widetilde{S}_t\right\}$ with finite state space $\widetilde{A}$. Let f be a probabilistic function from $\widetilde{A}$ into A such that $X_t = f\left(\widetilde{S}_t\right)$. In the case of a hidden Markov chain (Ephraim and Mehrav, 2002), the function f is such that $f\left(i\right) = f\left(j\right)$ may be satisfied for some different i, j. Here, we will restrict our attention to two-level stochastic processes $\left\{\widetilde{S}_t, X_t\right\}$ such that f is one-to-many. Hence, the state space corresponds to a partition of the output space. In this case, the Markov chain $\left\{\widetilde{S}_t\right\}$ is no longer 'hidden' since the value taken by X_t determines uniquely the value taken by $\widetilde{S}_t$; hence, $\widetilde{S}_t = \widetilde{s}_t$ if and only if $x_t \in \widetilde{A}\left(\widetilde{s}_t\right)$.

The differences between the two ways of constructing lumped processes may be summarized as follows:

- The function f is one-to-many while the function h is many-to-one. In the classical approach, the Markov chain $\{S_t\}$ is therefore 'hidden' while in the proposed approach, the Markov chain $\left\{\widetilde{S}_t\right\}$ is not hidden. Hence, $\{S_t, Y_t\}$ is not in general a Markov chain while $\left\{\widetilde{S}_t, X_t\right\}$ is always a Markov chain (see below for justifications).
- In the classical approach the objective is to determine if the lumped process $\{S_t, Y_t\}$ can be reparameterized as a more parsimonious lumped Markov chain $\left\{\widetilde{S}_t\right\}$, both defined on the coarser state space $\widetilde{A}$. In the proposed approach, the objective is to determine if the original Markov chain $\{S_t\}$ can be reparameterized as a more parsimonious lumped process $\left\{\widetilde{S}_t, X_t\right\}$, both defined on the original state space A.

- In the classical approach, the target process is the lumped Markov chain $\left\{\widetilde{S}_t\right\}$ defined on the coarser state space $\widetilde{A}$ while in the proposed approach, the target process is the two-level lumped process $\left\{\widetilde{S}_t, X_t\right\}$. Hence, in the classical approach, only results in the coarser state space are of interest (Kemeny and Snell, 1976) while in the proposed approach, results both in the coarser state space and in the original state space may be of interest.

Since f is a probabilistic function (and not a deterministic function), different conditional independence structures may be proposed to relate the output process to the state process. In this manner, the proposed approach will be constructive, the state process $\left\{\widetilde{S}_t\right\}$ being by definition a finite state Markov chain and different possible lumped processes $\left\{\widetilde{S}_t, X_t\right\}$ corresponding to different conditional independence assumptions will be investigated. To preserve the Markovian property within the lumped process $\left\{\widetilde{S}_t, X_t\right\}$, the output variable X_t cannot depend directly on (state or output) variables delayed from more than one time step, i.e. X_{t-r} or $\widetilde{S}_{t-r}$ with $r > 1$. Hence,

$$\begin{aligned} & P\left(X_t = x_t | X_0^{t-1} = x_0^{t-1}\right) \\ = \; & P\left(X_t = x_t, \widetilde{S}_t = \widetilde{s}_t | X_0^{t-1} = x_0^{t-1}, \widetilde{S}_0^{t-1} = \widetilde{s}_t^{t-1}\right) \\ = \; & P\left(X_t = x_t, \widetilde{S}_t = \widetilde{s}_t | X_{t-1} = x_{t-1}, \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right), \end{aligned}$$

where $X_0^{t-1} = x_0^{t-1}$ is a shorthand for $X_0 = x_0, \ldots, X_{t-1} = x_{t-1}$ (this convention transposes to the state sequence $\widetilde{S}_0^{t-1} = \widetilde{s}_t^{t-1}$). Then,

$$\begin{aligned} & P\left(X_t = x_t, \widetilde{S}_t = \widetilde{s}_t | X_{t-1} = x_{t-1}, \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) \\ = \; & P\left(X_t = x_t | \widetilde{S}_t = \widetilde{s}_t, X_{t-1} = x_{t-1}, \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) P\left(\widetilde{S}_t = \widetilde{s}_t | \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right), \end{aligned}$$

since $\left\{\widetilde{S}_t\right\}$ is a Markov chain.
Since f is one-to-many, we have

$$P\left(X_t = x_t | \widetilde{S}_t = \widetilde{s}_t, X_{t-1} = x_{t-1}, \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) = P\left(X_t = x_t | \widetilde{S}_t = \widetilde{s}_t, X_{t-1} = x_{t-1}\right),$$

i.e. once X_{t-1} is known, $\widetilde{S}_{t-1}$ conveys no further information about X_t. Hence, the output variable X_t which depends on $\widetilde{S}_t$ can also depend on either $\widetilde{S}_{t-1}$ or X_{t-1}. In the case of first-order Markov chains, three different types of lumped processes $\left\{\widetilde{S}_t, X_t\right\}$ (termed state-dependent lumped process, two-state-dependent lumped process and output-state-dependent lumped process) can then be defined

$$
\begin{aligned}
& P\left(X_t = x_t | X_0^{t-1} = x_0^{t-1}\right) \\
= & \begin{cases}
P\left(X_t = x_t | \widetilde{S}_t = \widetilde{s}_t\right) P\left(\widetilde{S}_t = \widetilde{s}_t | \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) & \text{state-dependent,} \\
P\left(X_t = x_t | \widetilde{S}_t = \widetilde{s}_t, \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) P\left(\widetilde{S}_t = \widetilde{s}_t | \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) & \text{two-state-dependent,} \\
P\left(X_t = x_t | \widetilde{S}_t = \widetilde{s}_t, X_{t-1} = x_{t-1}\right) P\left(\widetilde{S}_t = \widetilde{s}_t | \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) & \text{output-state-dependent.}
\end{cases}
\end{aligned}
$$

Hence, for a given partition $\widetilde{A}$ of A, the three classes of lumped processes are nested, output-state-dependent lumped processes containing two-state-dependent lumped processes, which themselves contain state-dependent lumped processes.
Since

$$
P\left(X_t = x_t | X_0^{t-1} = x_0^{t-1}\right) = \begin{cases}
P\left(X_t = x_t | \widetilde{S}_{t-1} = \widetilde{s}_{t-1}\right) & \text{(two-)state-dependent,} \\
P\left(X_t = x_t | X_{t-1} = x_{t-1}\right) & \text{output-state-dependent,}
\end{cases} \tag{3.1}
$$

the state-dependent and two-state-dependent lumped processes exhibit a stronger property than the Markovian property of the output-state-dependent lumped processes.

The conditional independence assumptions within a lumped process $\left\{\widetilde{S}_t, X_t\right\}$ can be expressed as a conditional independence graphs; see Lauritzen (1996) for an overview of graphical models and Smyth *et al.* (1997) for discussions of graphical models applied to hidden Markov chains and other two-level stochastic processes. In such a directed acyclic graph (Figure 1), each vertex represents a random variable, either a state variable $\widetilde{S}_t$ or an output variable X_t, at some time t. The absence of an arc (or directed edge) from a vertex Y_{t_1} (either of $\widetilde{S}$ or of X type) pointing towards a vertex Y_{t_2} $(t_1 \leq t_2)$ indicates that the two random variables Y_{t_1} and Y_{t_2} are conditionally independent given all the other random variables $(Y_t; t \leq t_2)$. The three conditional independence graphs corresponding to the three possible lumped processes are shown in Figure 1. Hence, arcs from vertices $\widetilde{S}_{t-1}$ pointing towards vertices X_t are added for two-state-dependent lumped processes (Figure 1b) while arcs from vertices X_{t-1} pointing towards vertices X_t are added for output-state-dependent lumped processes (Figure 1c) in comparison with state-dependent lumped processes (Figure 1a).

The stationary Markov chain $\left\{\widetilde{S}_t\right\}$ is uniquely determined by the transition probability matrix $\widetilde{P} = (\widetilde{p}_{ij};\ i, j = 0, \ldots, J-1)$.

State-dependent lumped process: The output process $\{X_t\}$ is related to the Markov chain $\left\{\widetilde{S}_t\right\}$ by the observation (or emission) probabilities

$$
b_j(n) = P\left(X_t = n | \widetilde{S}_t = j\right), \qquad j = 0, \ldots, J-1; n = 0, \ldots, N-1, \tag{3.2}
$$

with $\sum_{n \in \widetilde{A}(j)} b_j(n) = 1$.

The observation probabilities can be arranged as a $J \times N$ matrix, denoted by B, with all rows summing to one and each column containing exactly one non-zero entry (in the case of a

hidden Markov chain, each column contains at least one non-zero entry since the state-to-output mapping is such that $f(i) = f(j)$ may be satisfied for some different i, j).

Two-state-dependent lumped process: The output process $\{X_t\}$ is related to the Markov chain $\left\{\widetilde{S}_t\right\}$ by the following observation probabilities

$$b_{ij}(n) = P\left(X_t = n|\widetilde{S}_t = j, \widetilde{S}_{t-1} = i\right), \qquad i, j = 0, \ldots, J-1; n = 0, \ldots, N-1, \tag{3.3}$$

with $\sum_{n\in\widetilde{A}(j)} b_{ij}(n) = 1$.

Output-state-dependent lumped process: The output process $\{X_t\}$ is related to the Markov chain $\left\{\widetilde{S}_t\right\}$ by the following observation probabilities

$$b_{m,j}(n) = P\left(X_t = n|\widetilde{S}_t = j, X_{t-1} = m\right), \qquad \begin{array}{c} m = 0, \ldots, N-1; j = 0, \ldots, J-1; \\ n = 0, \ldots, N-1, \end{array} \tag{3.4}$$

with $\sum_{n\in\widetilde{A}(j)} b_{m,j}(n) = 1$.

In what follows, the notation p_{mn} will be used both for the transition probabilities of the original Markov chain $\{S_t\}$ - $p_{mn} = P(S_t = n|\ S_{t-1} = m)$ - and the output transition probabilities of the lumped process $\left\{\widetilde{S}_t, X_t\right\}$ - $p_{mn} = P(X_t = n|\ X_{t-1} = m)$ -. The objective is now to derive criteria to determine if an original Markov chain $\{S_t\}$ is lumpable with respect to a partition $\widetilde{A} = \left\{\widetilde{A}(0), \ldots, \widetilde{A}(J-1)\right\}$ of A, i.e. if the original Markov chain $\{S_t\}$ can be reparameterized as a more parsimonious lumped process $\left\{\widetilde{S}_t, X_t\right\}$. These criteria are direct consequences of properties of the output transition probabilities $p_{mn} = P(X_t = n|\ X_{t-1} = m)$ which are given below.

Proposition 1.
State-dependent lumped process: The output transition probabilities $p_{mn} = P(X_t = n|\ X_{t-1} = m)$ satisfy the two following criteria:
(i) row equality criterion

$$p_{mn} = b_j(n)\widetilde{p}_{ij} = p_{m'n}, \qquad m \neq m'; m, m' \in \widetilde{A}(i); n \in \widetilde{A}(j),$$

(ii) column ratio criterion

$$\frac{p_{mn}}{p_{mn'}} = \frac{b_j(n)\widetilde{p}_{ij}}{b_j(n')\widetilde{p}_{ij}} = \frac{b_j(n)}{b_j(n')}, \qquad m \in \widetilde{A}(i); n \neq n'; n, n' \in \widetilde{A}(j). \tag{3.5}$$

Hence $p_{mn}/p_{mn'}$ do not depend on m.

Two-state-dependent lumped process: The output transition probabilities $p_{mn} = P(X_t = n|\ X_{t-1} = m)$ satisfy the row equality criterion

$$p_{mn} = b_{ij}(n)\widetilde{p}_{ij} = p_{m'n}, \qquad m \neq m'; m, m' \in \widetilde{A}(i); n \in \widetilde{A}(j).$$

Output-state-dependent lumped process: The output transition probabilities $p_{mn} = P(X_t = n| X_{t-1} = m)$ satisfy the row sum criterion

$$\sum_{n\in\widetilde{A}(j)} p_{mn} = \widetilde{p}_{ij} = \sum_{n\in\widetilde{A}(j)} p_{m'n}, \qquad m \neq m'; m, m' \in \widetilde{A}(i). \tag{3.6}$$

Hence, in this case, we obtain the row sum criterion of the classical lumpability property which in our context is the weakest criterion.

These properties are direct consequences of the definition of lumped processes given by (3.2) (3.3) and (3.4).

Theorem 1.
(i) The original Markov chain $\{S_t\}$ can be reparameterized as a state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A if and only if the transition probabilities satisfy the row equality and the column ratio criteria.
(ii) The original Markov chain $\{S_t\}$ can be reparameterized as a two-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A if and only if the transition probabilities satisfy the row equality criterion.
(iii) The original Markov chain $\{S_t\}$ can be reparameterized as an output-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A if and only if the transition probabilities satisfy the row sum criterion.
This theorem follows directly from the properties of the output transition probabilities given in Proposition 1.

It is possible to adopt a different presentation of the criteria of reparameterization which puts the focus on the nested character of the three lumped processes for a given partition $\widetilde{A}$ of A. It is obvious that, for the three lumped processes, the output transition probabilities $p_{mn} = P(X_t = n| X_{t-1} = m)$ satisfy the row sum criterion (3.6). With reference to output-state-dependent lumped processes, it is necessary to add the following column ratio criterion for two-state-dependent lumped processes

$$\frac{p_{mn}}{p_{mn'}} = \frac{b_{ij}(n)\widetilde{p}_{ij}}{b_{ij}(n')\widetilde{p}_{ij}} = \frac{b_{ij}(n)}{b_{ij}(n')}, \qquad m \in \widetilde{A}(i); n \neq n'; n, n' \in \widetilde{A}(j). \tag{3.7}$$

Hence $p_{mn}/p_{mn'}$ do not depend on m for a fixed i.
And for state-dependent lumped processes, it is necessary to add the column ratio criterion (3.5). The row equality criterion is directly deduced from the row sum criterion and the most restricted column ratio criterion (3.7).

In the case of a state-dependent lumped process, the original transition probability matrix P is related to the transition probability matrix $\widetilde{P}$ and the observation probability matrix B of the lumped process $\left\{\widetilde{S}_t, X_t\right\}$ by

$$P = C\widetilde{P}B, \tag{3.8}$$

where C is the transpose of B such that all the non-zero entries are replaced by 1's (C is a $N \times J$ matrix whose j-th column is a vector consisting of 1's for the outputs aggregated in the j-th state and 0's otherwise). Since $BC = I$ (the j,jth entry is $\sum_{n\in\widetilde{A}(j)} b_j(n) = 1$ while the i,jth entry with $i \neq j$ is zero) where I denotes the $J \times J$ identity matrix, relation (3.8) can be rewritten as

$$BPC = \widetilde{P}.$$

This relation is used to compute $\widetilde{P}$ from P and C in the framework of the classical approach to lumpability (Kemeny and Snell, 1976). The matrix C is indeed the observation probability matrix of the lumped process $\{S_t, Y_t\}$ defined in the framework of the classical approach to lumpability; see Section 2. In this framework, the choice of B is by no means unique and all is needed is that the j-th row have non-zero entries for outputs aggregated in state j.

To illustrate the construction of the three types of lumped processes $\left\{\widetilde{S}_t, X_t\right\}$, consider an original 3-state Markov chain $\{S_t\}$ with transition probabilities $(p_{mn};\ m,n = 0,1,2)$. Suppose this Markov chain is lumpable with respect to the partition

$$\widetilde{A} = \{\{0\},\{1,2\}\} = \left\{\widetilde{A}(0), \widetilde{A}(1)\right\}.$$

Illustration with a state-dependent lumped process: The transition probabilities $(p_{mn};\ m,n = 0,1,2)$ of the original Markov chain $\{S_t\}$ are related to the transition probabilities $(\widetilde{p}_{ij};\ i,j = 0,1)$ and the observation probabilities $(b_j(n);\ j = 0,1;\ n = 0,1,2)$ of the state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ by

$$\begin{aligned} P &= C\widetilde{P}B \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \widetilde{p}_{00} & \widetilde{p}_{01} \\ \widetilde{p}_{10} & \widetilde{p}_{11} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & b_1(1) & b_1(2) \end{pmatrix} \\ &= \begin{pmatrix} \widetilde{p}_{00} & \widetilde{p}_{01}b_1(1) & \widetilde{p}_{01}b_1(2) \\ \widetilde{p}_{10} & \widetilde{p}_{11}b_1(1) & \widetilde{p}_{11}b_1(2) \\ \widetilde{p}_{10} & \widetilde{p}_{11}b_1(1) & \widetilde{p}_{11}b_1(2) \end{pmatrix}. \end{aligned}$$

The multiplication by the matrix C duplicates the rows corresponding to the states with more than one possible output and thus generates the row equality criterion while the multiplication by the matrix B adds columns with identical transition probabilities $\widetilde{p}_{ij}$ and observation probabilities constrained by $\sum_{n\in\widetilde{A}(j)} b_j(n) = 1$ for the outputs aggregated in a given state and thus generates the column ratio criterion.
Hence,

$$p_{1n} = p_{2n}, \qquad n = 0,1,2,$$

and

$$\frac{p_{01}}{p_{02}} = \frac{p_{11}}{p_{12}} = \frac{p_{21}}{p_{22}}, \qquad \text{where} \quad \frac{p_{11}}{p_{12}} \equiv \frac{p_{21}}{p_{22}}$$

($\equiv$ means both numerators and denominators equal).

The number of free parameters of the state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ is thus $N(N-1)$ subject to

$$p_{mn} = p_{m'n}, \qquad m \neq m'; m, m' \in \widetilde{A}(i); i = 0, \ldots, J-1; n = 0, \ldots, N-1,$$

$(N-J)(N-1)$ constraints (row equality criterion), and

$$\frac{p_{mn}}{p_{mn'}} = \frac{b_j(n)}{b_j(n')}, \qquad m = 0, \ldots, N-1; n \neq n'; n, n' \in \widetilde{A}(j); j = 0, \ldots, J-1,$$

$(N-1)(N-J)$ constraints (column ratio criterion). We should also take into account the $(N-J)(N-J)$ double constraints resulting from both the row equality and the column ratio criteria (see the above example). Finally, the number of free parameters is

$$\begin{aligned} & N(N-1) - (N-J)(N-1) - (N-1)(N-J) + (N-J)(N-J) \\ = \;& J(J-1) + N - J. \end{aligned} \tag{3.9}$$

In expression (3.9), $J(J-1)$ is the number of independent transition probabilities while $N-J$ is the number of independent observation probabilities.

Illustration with a *two-state-dependent lumped process*: The transition probabilities $(p_{mn};\ m, n = 0, 1, 2)$ of the original Markov chain $\{S_t\}$ are related to the transition probabilities $(\widetilde{p}_{ij};\ i, j = 0, 1)$ and the observation probabilities $(b_{ij}(n);\ i, j = 0, 1;\ n = 0, 1, 2)$ of the two-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ by

$$P = \begin{pmatrix} \widetilde{p}_{00} & \widetilde{p}_{01} b_{01}(1) & \widetilde{p}_{01} b_{01}(2) \\ \widetilde{p}_{10} & \widetilde{p}_{11} b_{11}(1) & \widetilde{p}_{11} b_{11}(2) \\ \widetilde{p}_{10} & \widetilde{p}_{11} b_{11}(1) & \widetilde{p}_{11} b_{11}(2) \end{pmatrix}.$$

Hence,

$$p_{1n} = p_{2n}, \qquad n = 0, 1, 2.$$

The number of free parameters of the lumped process $\left\{\widetilde{S}_t, X_t\right\}$ is thus $N(N-1)$ subject to $(N-J)(N-1)$ constraints (row equality criterion). Finally, the number of free parameters is

$$\begin{aligned} N(N-1) - (N-J)(N-1) &= J(N-1) \\ &= J(J-1) + J(N-J). \end{aligned}$$

Illustration with an output-state-dependent lumped process: The transition probabilities $(p_{mn}; m,n=0,1,2)$ of the original Markov chain $\{S_t\}$ are related to the transition probabilities $(\widetilde{p}_{ij}; i,j=0,1)$ and the observation probabilities $(b_{m,j}(n);\ m=0,1,2; j=0,1; n=0,1,2)$ of the output-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ by

$$P=\begin{pmatrix} \widetilde{p}_{00} & \widetilde{p}_{01}b_{0,1}(1) & \widetilde{p}_{01}b_{0,1}(2) \\ \widetilde{p}_{10} & \widetilde{p}_{11}b_{1,1}(1) & \widetilde{p}_{11}b_{1,1}(2) \\ \widetilde{p}_{10} & \widetilde{p}_{11}b_{2,1}(1) & \widetilde{p}_{11}b_{2,1}(2) \end{pmatrix}.$$

Hence,

$$p_{10}=p_{20}, \quad p_{11}+p_{12}=p_{21}+p_{22}.$$

The number of free parameters of the lumped process $\left\{\widetilde{S}_t, X_t\right\}$ is thus $N(N-1)$ subject to

$$\sum_{n\in A(j)} p_{mn}=\sum_{n\in A(j)} p_{m'n}, \qquad m\neq m'; m,m'\in \widetilde{A}(i); i,j=0,\ldots,J-1,$$

$(N-J)(J-1)$ constraints (row sum criterion); see Thomas and Barr (1977) in the framework of the classical approach to lumpability. Finally, the number of free parameters is

$$N(N-1)-(N-J)(J-1)=J(J-1)+N(N-J).$$

It is interesting to illustrate with this example the alternative point of view which puts the focus on the nested character of the three lumped processes for a given partition $\widetilde{A}$ of A. The number of free parameters of the lumped process $\left\{\widetilde{S}_t, X_t\right\}$ is $N(N-1)$ subject to $(N-J)(J-1)$ constraints (row sum criterion). In the case of the two-state-dependent lumped process, $(N-J)(N-J)$ constraints corresponding to the column ratio criterion (3.7) are added, and in the case of the state-dependent lumped process, $(N-1)(N-J)$ constraints corresponding to the column ratio criterion (3.5). In the example, to the constraints

$$p_{10}=p_{20}, \quad p_{11}+p_{12}=p_{21}+p_{22},$$

the constraint

$$\frac{p_{11}}{p_{12}}=\frac{p_{21}}{p_{22}}$$

is added in the case of the two-state-dependent lumped process, and the constraints

$$\frac{p_{01}}{p_{02}}=\frac{p_{11}}{p_{12}}=\frac{p_{21}}{p_{22}}$$

are added in the case of the state-dependent lumped process. It should be noted that the row sum criterion (3.6) and the column ratio criteria (3.5) or (3.7) are non-redundant in the sense that they do not generate double constraints. The number of free parameters of the lumped process $\left\{\widetilde{S}_t, X_t\right\}$ is thus

$$N(N-1)-(N-J)(J-1)-\begin{cases}(N-1)(N-J) & \text{state-dependent,}\\ (N-J)(N-J) & \text{two-state-dependent,}\\ & \text{output-state-dependent.}\end{cases}$$

In the case of an original first-order Markov chain, one way to relate the three types of lumped processes is by considering their respective conditional independence graphs presented in Figure 1. A complementary perspective is given by comparing their respective numbers of free parameters as a function of the number of states J for a fixed number of outputs N

$$\begin{aligned}\varphi_N(J) &= J(J-1)+N-J && \text{state-dependent,}\\ \phi_N(J) &= J(J-1)+J(N-J)=J(N-1) && \text{two-state-dependent,}\\ \psi_N(J) &= J(J-1)+N(N-J) && \text{output-state-dependent.}\end{aligned}$$

These three functions are represented in Figure 2 for $N=11$. It can be seen that both $\varphi_N(J)$ and $\phi_N(J)$ are strictly increasing functions from $\varphi_N(1)=\phi_N(1)=N-1$ (zero-order Markov chain) to $\varphi_N(N)=\phi_N(N)=N(N-1)$ (first-order Markov chain). The behaviour of $\psi_N(J)$ is very different since $\psi_N(J)=\psi_N(N-J+1)$ and $\psi_N(1)=\psi_N(N)=N(N-1)$ (first-order Markov chain). The case $J=1$ can be interpreted as an N-state first-order Markov chain built on the successive random variables X_t while the case $J=N$ can be interpreted as an N-state first-order Markov chain built on the successive random variables $\widetilde{S}_t$ (this behaviour is naturally expressed in the conditional independence graph in Figure 1c). In the case where N is odd, the function $\psi_N(J)$ reaches its minimum for $J=(N+1)/2$ where $\psi_N(J)=(N-1)(3N+1)/4$. The maximum gain in number of free parameters, which is $(N-1)^2/4$, seems rather limited. Moreover, the symmetric shape of $\psi_N(J)$ appears to be counterintuitive regarding the concept of lumpability where a strictly increasing function such as $\varphi_N(J)$ or $\phi_N(J)$ is expected. Hence, for output-state-dependent lumped processes, a recommendation would be to evaluate only partitions $\widetilde{A}$ of A such that $J \geq N/2$.

Since in the case of first-order Markov chains, memories coincide with states, models built by lumping states on the basis of the row equality criterion can be seen as variable-order Markov chains (Weinberger *et al.*, 1995; Ron *et al.*, 1996; Bühlmann and Wyner, 1999). Hence, the (two)-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ is partly a first-order Markov chain (with $J(J-1)$ free parameters corresponding to the first-order part) and partly a zero-order Markov chain (with $(N-J)$ free parameters for the state-dependent lumped process and $J(N-J)$ free parameters for the two-state-dependent lumped process corresponding to the zero-order part); this is also a direct consequence of (3.1). The specificity of state-dependent lumped processes lies in the reversibility property which is described in the proposition below.

Proposition 2.
If the original Markov chain $\{S_t\}$ can be reparameterized as a state-dependent lumped process

$\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A, then the reversed chain (obtained by reversing the direction of time) can be reparameterized as a state-dependent lumped process with respect to the same partition $\widetilde{A}$ of A.

Proof: The transition probabilities $(p_{mn};\ m,n=0,\ldots,N-1)$ of the original Markov chain satisfy the row equality criterion

$$p_{mn}=p_{m'n}, \qquad m\neq m';m,m'\in\widetilde{A}\,(i)\,;n\in\widetilde{A}\,(j)\,,$$

and the column ratio criterion

$$\frac{p_{nm}}{p_{nm'}}=\frac{b_i(m)}{b_i(m')}, \qquad n\in\widetilde{A}\,(j)\,;m\neq m';m,m'\in\widetilde{A}\,(i)\,.$$

Since the process is stationary, we have

$$\frac{p_{nm}}{p_{nm'}}=\frac{\pi_m}{\pi_{m'}},$$

where $\pi_m,\pi_{m'}$ are the stationary probabilities for the states m and m'.

The stationary reversed Markov chain is uniquely determined by the transition probability matrix $Q=(q_{mn};\ m,n=0,\ldots,N-1)$ with (Feller, 1968; Kulkarni,1995)

$$q_{mn}=\frac{p_{nm}\pi_n}{\pi_m}.$$

The column ratio criterion can be rewritten as

$$\frac{p_{nm}\pi_n}{\pi_m}=\frac{p_{nm'}\pi_n}{\pi_{m'}}.$$

Hence $q_{mn}=q_{m'n}$.
The row equality criterion can be rewritten as

$$\frac{p_{mn}}{\pi_n}=\frac{p_{m'n}}{\pi_n}.$$

Hence,

$$\frac{q_{nm}}{q_{nm'}}=\frac{\pi_m}{\pi_{m'}}. \quad \square$$

4 Model selection methods

In the following, the model selection methods which are direct transpositions of the methods used for the determination of the order of a Markov chain will be presented for state-dependent lumped processes, the cases of two-state-dependent and output-state-dependent lumped processes being outlined at the end of this Section. Let us denote H_0 the hypothesis that $\{S_t\}$ can be reparameterized as a state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A. Note that H_0 is contained in the larger hypothesis H_1 corresponding to the fact that $\{S_t\}$ is a first-order Markov chain. A procedure for testing such hypotheses can be based on a χ^2 or a

likelihood ratio test. Let f_{mn} denote the observed frequency of pairs of outputs in which output m is followed by output n. For the corresponding marginal frequencies, we use the customary dot notation

$$f_{m\cdot} = \sum_n f_{mn}, \qquad f_{\cdot n} = \sum_m f_{mn}.$$

A similar definition applies to f_{ij}, $f_{i\cdot}$ and $f_{\cdot j}$ at the state level. We have the obvious following relationships for each state j

$$f_{\cdot j} = \sum_{n \in \widetilde{A}(j)} f_{\cdot n}.$$

One simple way for testing a lumpability hypothesis is to calculate the expected frequencies of different pairs of successive outputs, assuming a given partition of the original state space, and to compare these expected frequencies with the observed frequencies f_{mn}. By a direct transposition of Billingsley (1961), the χ^2 statistic

$$\sum_{m,n} \frac{\left(f_{mn} - f_{m\cdot}\,\widehat{b}_j(n)\,\widehat{p}_{ij}\right)^2}{f_{m\cdot}\,\widehat{b}_j(n)\,\widehat{p}_{ij}}, \qquad m \in \widetilde{A}(i)\,; n \in \widetilde{A}(j)\,, \tag{4.1}$$

where $\widehat{p}_{ij}$ and $\widehat{b}_j(n)$ are the estimates of $\widetilde{p}_{ij}$ and $b_j(n)$ respectively

$$\widehat{p}_{ij} = \frac{f_{ij}}{f_{i\cdot}}, \quad \widehat{b}_j(n) = \frac{f_{\cdot n}}{f_{\cdot j}},$$

has asymptotically a χ^2 distribution with $N(N-1) - \{J(J-1) + N - J\} = (N-J)(N+J-2)$ degrees of freedom. These degrees of freedom are maximal for $J = 1$ which corresponds to the classical χ^2 test with $(N-1)^2$ degrees of freedom for testing the zero-order hypothesis (purely random sequence) within the first-order hypothesis (Billingsley, 1961). The zero-order Markov chain can be seen as a state-dependent lumped process with a single state.

Likelihood ratio tests can also be used to test lumpability hypotheses. The likelihood of the output sequence $x_0^{\tau-1}$ which uniquely determines the state sequence $\widetilde{s}_0^{\tau-1}$ is given by

$$\begin{aligned} & P\left(X_0 = x_0, \widetilde{S}_0 = \widetilde{s}_0, \ldots, X_{\tau-1} = x_{\tau-1}, \widetilde{S}_{\tau-1} = \widetilde{s}_{\tau-1}; \widetilde{P}, B\right) \\ = \;& \widetilde{\pi}_{\widetilde{s}_0} b_{\widetilde{s}_0}(x_0) \prod_{t=1}^{\tau-1} \widetilde{p}_{\widetilde{s}_{t-1}\widetilde{s}_t} b_{\widetilde{s}_t}(x_t) \\ = \;& \widetilde{\pi}_{\widetilde{s}_0} b_{\widetilde{s}_0}(x_0)\, L\left(\widetilde{P}, B\right). \end{aligned}$$

where $\widetilde{\pi}_{\widetilde{s}_0}$ is the stationary probability of state $\widetilde{s}_0$. The likelihoods under the two models, the original one and the lumped one, are essentially

$$
\begin{aligned}
L(P) &= \prod_{m,n} p_{mn}^{f_{mn}}, \\
L\left(\widetilde{P}, B\right) &= \prod_{i,j} \widetilde{p}_{ij}^{f_{ij}} \left\{ \prod_{j} \prod_{n \in \widetilde{A}(j)} b_j(n)^{f_{\cdot n}} \right\}.
\end{aligned}
$$

We must compute the maxima of the log-likelihoods under the two models. We already know that $\widehat{p}_{mn} = f_{mn}/f_{m\cdot}$, $\widehat{p}_{ij} = f_{ij}/f_{i\cdot}$ and $\widehat{b}_j(n) = f_{\cdot n}/f_{\cdot j}$. The log-likelihood of the output sequence $x_0^{\tau-1}$ for the estimated parameters $\widehat{P}, \widehat{B}$ is given by

$$
\log L\left(\widehat{P}, \widehat{B}\right) = \sum_{i,j} f_{ij} \log \frac{f_{ij}}{f_{i\cdot}} + \sum_{j} \sum_{n \in \widetilde{A}(j)} f_{\cdot n} \log \frac{f_{\cdot n}}{f_{\cdot j}}. \tag{4.2}
$$

The log-likelihood ratio statistic for testing H_0 is

$$
\begin{aligned}
& 2\left\{ \sum_{m,n} f_{mn} \log \frac{f_{mn}}{f_{m\cdot}} - \left(\sum_{i,j} f_{ij} \log \frac{f_{ij}}{f_{i\cdot}} + \sum_{j} \sum_{n \in \widetilde{A}(j)} f_{\cdot n} \log \frac{f_{\cdot n}}{f_{\cdot j}} \right) \right\} \\
= \; & 2 \sum_{m,n} f_{mn} \log \frac{f_{mn}/f_{m\cdot}}{\left(f_{ij}/f_{i\cdot}\right)\left(f_{\cdot n}/f_{\cdot j}\right)} \qquad m \in \widetilde{A}(i)\,; n \in \widetilde{A}(j) \\
= \; & 2 \sum_{m,n} f_{mn} \log \frac{f_{mn}}{f_{m\cdot}\, \widehat{p}_{ij} \widehat{b}_j(n)},
\end{aligned} \tag{4.3}
$$

which asymptotically has a χ^2 distribution with $N(N-1) - \{J(J-1) + N - J\}$ degrees of freedom.

The χ^2 (4.1) and the log-likelihood ratio (4.3) statistics are related under the null hypothesis by the well-known asymptotic equivalence

$$
\sum_{m,n} \frac{\left(f_{mn} - f_{m\cdot}\, \widehat{b}_j(n)\, \widehat{p}_{ij}\right)^2}{f_{m\cdot}\, \widehat{b}_j(n)\, \widehat{p}_{ij}} \simeq 2 \sum_{m,n} f_{mn} \log \frac{f_{mn}}{f_{m\cdot}\, \widehat{p}_{ij} \widehat{b}_j(n)}, \qquad m \in \widetilde{A}(i)\,; n \in \widetilde{A}(j)\,.
$$

If different lumpability hypotheses are to be evaluated, a first possibility is to apply a collection of either χ^2 or likelihood ratio tests. But, this approach is not recommended since there are no formal rules that rigorously define how the various P-values might be used to select a final model; see Akaike (1974), Kass and Raftery (1995) and Burnham and Anderson (2002). Moreover, the fact that χ^2 or likelihood ratio tests exist only for nested models may be too restrictive. Hence, hypothesis testing should only be used as a simple test in the case where a single lumpability hypothesis is considered. Another possibility is to use penalized log-likelihood criteria such as Akaike's Information Criterion (AIC) (Akaike, 1974; Burnham and Anderson, 2002) or Bayesian Information Criterion (BIC) (Schwarz, 1978; Katz, 1981; Kass and Raftery, 1995; Csiszár and Shields, 2000) to evaluate lumpability hypotheses. A substantial advantage of penalized log-likelihood criteria over hypothesis testing is that they are also valid for nonnested

models (Kass and Raftery, 1995; Burnham and Anderson, 2002). For example, in the BIC case, the maximum log-likelihood for the first-order Markov chain $\{S_t\}$ is penalized by a function of both the number of free parameters and the sample size τ

$$2\sum_{m,n} f_{mn} \log \frac{f_{mn}}{f_{m\cdot}} - N\left(N-1\right) \log \tau. \tag{4.4}$$

The maximum log-likelihood for the state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ given by (4.2) is similarly penalized

$$2\left(\sum_{i,j} f_{ij} \log \frac{f_{ij}}{f_{i\cdot}} + \sum_{j} \sum_{n \in \widetilde{A}(j)} f_{\cdot n} \log \frac{f_{\cdot n}}{f_{\cdot j}}\right) - \left\{J\left(J-1\right) + N - J\right\} \log \tau. \tag{4.5}$$

The BIC favors the model which gives the maximum penalized log-likelihood between (4.4) and (4.5). The above procedure can be directly extended to multiple, either nested or nonnested, lumpability hypotheses (corresponding in this latter case to different nonnested partitions of A).

The model selection methods presented above can be directly transposed to two-state-dependent lumped processes by remarking that the estimate of $b_{ij}\left(n\right)$ is given by

$$\widehat{b}_{ij}\left(n\right) = \frac{\sum_{u \in \widetilde{A}(i)} f_{un}}{\sum_{u \in \widetilde{A}(i), v \in \widetilde{A}(j)} f_{uv}} = \frac{\sum_{u \in \widetilde{A}(i)} f_{un}}{f_{ij}},$$

and that the log-likelihood of the output sequence $x_0^{\tau-1}$ for the estimated parameters $\widehat{P}, \widehat{B}$ is given by

$$\log L\left(\widehat{P}, \widehat{B}\right) = \sum_{i,j} f_{ij} \log \frac{f_{ij}}{f_{i\cdot}} + \sum_{i,j} \sum_{n \in \widetilde{A}(j)} \left(\sum_{u \in \widetilde{A}(i)} f_{un}\right) \log \frac{\sum_{u \in \widetilde{A}(i)} f_{un}}{f_{ij}}.$$

The model selection methods presented above can also be directly transposed to output-state-dependent lumped processes. For instance, the χ^2 statistic

$$\sum_{m,n} \frac{\left(f_{mn} - f_{m\cdot}\, \widehat{b}_{m,j}\left(n\right) \widehat{p}_{ij}\right)^2}{f_{m\cdot}\, \widehat{b}_{m,j}\left(n\right) \widehat{p}_{ij}}, \qquad m \in \widetilde{A}\left(i\right); n \in \widetilde{A}\left(j\right),$$

where $\widehat{p}_{ij}$ and $\widehat{b}_{m,j}\left(n\right)$ are the estimates of $\widetilde{p}_{ij}$ and $b_{m,j}\left(n\right)$ respectively

$$\widehat{p}_{ij} = \frac{f_{ij}}{f_{i\cdot}}, \quad \widehat{b}_{m,j}\left(n\right) = \frac{f_{mn}}{\sum_{v \in \widetilde{A}(j)} f_{mv}},$$

has asymptotically a χ^2 distribution with $N\left(N-1\right) - \left\{J\left(J-1\right) + N\left(N-J\right)\right\} = \left(N-J\right)\left(J-1\right)$ degrees of freedom (rather than $N^2 - \left\{J\left(J-1\right) + N\left(N-J\right)\right\}$ proposed by Thomas and Barr (1977)). These degrees of freedom are maximal for $J = \left(N+1\right)/2$ (see Section 3).
It can be noted that

$$\widehat{p}_{ij}\widehat{b}_{m,j}\left(n\right)=\left(\frac{\sum_{u\in\widetilde{A}(i),v\in\widetilde{A}(j)}f_{uv}}{\sum_{u\in\widetilde{A}(i)}f_{u\cdot}}\right)\frac{f_{mn}}{\sum_{v\in\widetilde{A}(j)}f_{mv}}$$

is the maximum likelihood estimator of p_{mn} under the constraints (2.1) proposed by Thomas and Barr (1977). In fact, the test proposed by Thomas and Barr is the χ^2 test for testing an output-state-dependent lumpability hypothesis within the general first-order hypothesis.

Using the log-likelihood of the output sequence $x_0^{\tau-1}$ for the estimated parameters $\widehat{P},\widehat{B}$ given by

$$\log L\left(\widehat{P},\widehat{B}\right)=\sum_{i,j}f_{ij}\log\frac{f_{ij}}{f_{i\cdot}}+\sum_{m}\sum_{j}\sum_{n\in\widetilde{A}(j)}f_{mn}\log\frac{f_{mn}}{\sum_{v\in\widetilde{A}(j)}f_{mv}},$$

the likelihood ratio test or the penalized log-likelihood criteria (AIC or BIC) can directly be derived.

5 Lumped processes based on high-order Markov chains

The construction of lumped processes based on first-order Markov chains can be transposed to higher-order Markov chains. For instance, if the original process $\{S_t\}$ is a stationary second-order Markov chain with transition probability matrix $P=(p_{kmn};\ k,m,n=0,\ldots,N-1)$, the corresponding lumped process $\left\{\widetilde{S}_t,X_t\right\}$ is constructed from a stationary second-order Markov chain with transition probability matrix $\widetilde{P}=(\widetilde{p}_{hij};\ h,i,j=0,\ldots,J-1)$ while the definition of the observation probabilities remains unchanged with respect to the first-order case; see (3.2), (3.3) and (3.4). In what follows, we assume that the transition probabilities p_{kmn} (respectively $\widetilde{p}_{hij}$) are arranged as a $N^2\times N$ (respectively $J^2\times J$) matrix with all rows, each row corresponding to a given memory, summing to one. The conditional independence assumptions within the lumped processes $\left\{\widetilde{S}_t,X_t\right\}$ can be translated into directed acyclic graphs which are directly deduced from the graphs in Figure 1 by adding arcs from vertices $\widetilde{S}_{t-2}$ pointing towards vertices $\widetilde{S}_t$. High-order Markov chains can be seen as first-order Markov chains defined on an enlarged state space corresponding to the possible memories. Hence, first-order Markov chain statistical theory applies to higher-order Markov chains (Billingsley, 1961; Guttorp 1995) and the model selection methods presented in Section 4 for lumped processes based on first-order Markov chains can directly be transposed to lumped processes based on higher-order Markov chains. We can now outline the main steps in the derivation of the transposition of the results presented in Section 3 for first-order Markov chains.

Proposition 3.
State-dependent lumped process: The output transition probabilities $p_{kmn}=P\left(X_t=n|\ X_{t-1}=m,\ X_{t-2}=k\right)$ satisfy the two following criteria:
(i) row equality criterion

$$p_{kmn} = b_j(n)\widetilde{p}_{hij} = p_{k'm'n}, \qquad k \neq k' \text{ or } m \neq m'; k, k' \in \widetilde{A}(h); m, m' \in \widetilde{A}(i); n \in \widetilde{A}(j),$$

(ii) column ratio criterion

$$\frac{p_{kmn}}{p_{kmn'}} = \frac{b_j(n)\widetilde{p}_{hij}}{b_j(n')\widetilde{p}_{hij}} = \frac{b_j(n)}{b_j(n')}, \qquad k \in \widetilde{A}(h); m \in \widetilde{A}(i); n \neq n'; n, n' \in \widetilde{A}(j). \tag{5.1}$$

Hence $p_{kmn}/p_{kmn'}$ do not depend on both k and m.

Two-state-dependent lumped process: The output transition probabilities $p_{kmn} = P(X_t = n| X_{t-1} = m, X_{t-2} = k)$ satisfy the two following criteria:
(i) row equality criterion

$$p_{kmn} = b_{ij}(n)\widetilde{p}_{hij} = p_{k'm'n}, \qquad k \neq k' \text{ or } m \neq m'; k, k' \in \widetilde{A}(h); m, m' \in \widetilde{A}(i); n \in \widetilde{A}(j),$$

(ii) column ratio criterion

$$\frac{p_{kmn}}{p_{kmn'}} = \frac{b_{ij}(n)\widetilde{p}_{hij}}{b_{ij}(n')\widetilde{p}_{hij}} = \frac{b_{ij}(n)}{b_{ij}(n')}, \qquad k \in \widetilde{A}(h); m \in \widetilde{A}(i); n \neq n'; n, n' \in \widetilde{A}(j). \tag{5.2}$$

Hence $p_{kmn}/p_{kmn'}$ do not depend on both k and m for a fixed i. We have therefore sub-column ratio criteria that correspond to each possible preceding state i.

Output-state-dependent lumped process: The output transition probabilities $p_{kmn} = P(X_t = n| X_{t-1} = m, X_{t-2} = k)$ satisfy the two following criteria:
(i) row sum criterion

$$\sum_{n \in \widetilde{A}(j)} p_{kmn} = \widetilde{p}_{hij} = \sum_{n \in \widetilde{A}(j)} p_{k'm'n}, \qquad k \neq k' \text{ or } m \neq m'; k, k' \in \widetilde{A}(h); m, m' \in \widetilde{A}(i), \tag{5.3}$$

(ii) column ratio criterion

$$\frac{p_{kmn}}{p_{kmn'}} = \frac{b_{m,j}(n)\widetilde{p}_{hij}}{b_{m,j}(n')\widetilde{p}_{hij}} = \frac{b_{m,j}(n)}{b_{m,j}(n')}, \qquad k \in \widetilde{A}(h); m \in \widetilde{A}(i); n \neq n'; n, n' \in \widetilde{A}(j). \tag{5.4}$$

Hence $p_{kmn}/p_{kmn'}$ do not depend on k. In fact, we have sub-column ratio criteria that correspond to each possible preceding output m.

Corollary 1. The output transition probabilities $p_{kmn} = P(X_t = n| X_{t-1} = m, X_{t-2} = k)$ of an output-state-dependent lumped process satisfy the following row equality criterion

$$p_{kmn} = b_{m,j}(n)\widetilde{p}_{hij} = p_{k'mn}, \qquad k \neq k'; k, k' \in \widetilde{A}(h); m \in \widetilde{A}(i); n \in \widetilde{A}(j).$$

Proof. Suppose that $n \in \widetilde{A}(j)$,

$$\begin{aligned} \sum_{v \in \widetilde{A}(j)} p_{kmv} &= p_{kmn}\left(1 + \frac{\sum_{v \neq n, v \in \widetilde{A}(j)} b_{m,j}(v)}{b_{m,j}(n)}\right) \\ &= \frac{p_{kmn}}{b_{m,j}(n)}. \end{aligned}$$

Applying the row sum criterion (5.3) for a fixed m completes the proof. □

Theorem 2.
(i) The original second-order Markov chain $\{S_t\}$ can be reparameterized as a state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A if and only if the transition probabilities satisfy the row equality criterion and the column ratio criterion defined in (5.1).
(ii) The original second-order Markov chain $\{S_t\}$ can be reparameterized as a two-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A if and only if the transition probabilities satisfy the row equality criterion and the column ratio criterion defined in (5.2).
(iii) The original second-order Markov chain $\{S_t\}$ can be reparameterized as an output-state-dependent lumped process $\left\{\widetilde{S}_t, X_t\right\}$ with respect to the partition $\widetilde{A}$ of A if and only if the transition probabilities satisfy the the row sum criterion and the column ratio criterion defined in (5.4).

Consider now lumped processes based on rth-order Markov chains and assume that the transition probabilities of the original rth-order Markov chain $\{S_t\}$ (respectively the rth-order Markov chain $\left\{\widetilde{S}_t\right\}$) are arranged as a $N^r \times N$ (respectively $J^r \times J$) matrix with all rows, each row corresponding to a given memory, summing to one. Another way of generalization to high-order Markov chains consists of linking the observation probabilities to the order of both the state process $\left\{\widetilde{S}_t\right\}$ and the original Markov chain $\{S_t\}$ instead of leaving them fixed and independent of the order. Recall that to preserve the rth-order Markovian property within the lumped process $\left\{\widetilde{S}_t, X_t\right\}$, the output variable X_t cannot depend directly on (state or output) variables delayed from more than r time steps. More precisely, the number of free parameters of the $(r+1)$-state-dependent lumped process and the r-output-state-dependent lumped process are then

$$\begin{aligned} &J^r(J-1) + J^r(N-J) = J^r(N-1) && (r+1)\text{-state-dependent}, \\ &J^r(J-1) + N^r(N-J) && r\text{-output-state-dependent}. \end{aligned}$$

The corresponding conditional independence graphs in the case of second-order Markov chains

are shown in Figure 3. For rth-order Markov chains with $r > 2$, it is possible to define many classes of lumped processes intermediate between the lumped processes where the observation probabilities are defined as for lumped processes based on first-order Markov chains (see Section 3) and lumped processes with linked observation probabilities as discussed above.

For an original rth-order Markov chain $\{S_t\}$, the number of constraints on the $N^r(N-1)$ independent transition probabilities is

row sum criterion:

$(N^r - J^r)(J-1)$.

row equality criteria:

$(N^r - J^r)(N-1)$	k-state-dependent lumped process $(1 \leq k \leq r+1)$,
$N^k\left(N^{r-k} - J^{r-k}\right)(N-1)$	k-output-state-dependent lumped process $(1 \leq k \leq r)$.

column ratio criteria:

$\left(N^r - J^{k-1}\right)(N-J)$	k-state-dependent lumped process $(1 \leq k \leq r+1)$,
$\left(N^r - N^k\right)(N-J)$	k-output-state-dependent lumped process $(1 \leq k \leq r)$,

double constraints resulting from the row equality and column ratio criteria:

$(N^r - J^r)(N-J)$	k-state-dependent lumped process $(1 \leq k \leq r+1)$.

The row sum criterion constitutes a base condition for lumpability and defines the maximum number of free parameters of a lumped process for a fixed order and a given partition of the original state space (or the minimum gain in number of free parameters with reference to an original rth-order Markov chain). With reference to this r-output-state-dependent lumped process, many more parsimonious lumped processes can be defined. The combination of the row sum criterion and the appropriate column ratio criterion constitutes a general way for stating conditions of reparameterization of an original high-order Markov chain. In the case of no direct dependencies between outputs (k-state-dependent lumped processes with $1 \leq k \leq r+1$), the combination of the row equality criterion with the appropriate column ratio criterion (which are partly redundant) constitutes a more intuitive way for stating conditions of reparameterization. In the case of direct dependencies between outputs (k-output-state-dependent lumped processes with $1 \leq k \leq r$), the row equality criterion is simply a byproduct of the row sum criterion and the appropriate column ratio criterion (see Corollary 1) and cannot be used in conjunction with solely this column ratio criterion to define conditions of reparameterization.

6 Application to DNA sequence analysis

Our objective here is to assess the relevance of lumped processes that incorporate knowledge relative to nucleotide grouping for the exploratory analysis of DNA sequences.

A deoxyribonucleic acid (DNA) molecule consists of two complementary sequences of nucleotides that twist to form a double helix. Each sequence is composed of four nucleotides consisting of a deoxyribose sugar, a phosphate group and a purine or pyrimidine base. Purine bases are adenine (A) and guanine (G); pyrimidine bases are cytosine (C) and thymine (T). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with

the 5' carbon of one sugar linked to the 3' carbon of the next, giving the sequence an orientation. The two single strands of a DNA molecule are connected by hydrogen bonding between complementary bases. Guanine pairs specifically with cytosine, forming three hydrogen bonds, and adenine pairs with thymine forming two hydrogen bonds. Hence, it may be useful to take into account some grouping of states relative either to the purine-pyrimidine (A/G, C/T), to the strong-weak hydrogen bonding (G/C, A/T) or to the keto-amino (T/G, A/C) classification in the analysis of DNA sequences.

Raftery and Tavaré (1994) were the first to test lumpability hypotheses in DNA sequences. For analysing mouse gene introns, they adopted the following procedure. In a first stage, they applied the Thomas and Barr (1977) method, based on a first-order Markov chain assumption, to test if some states may be lumped. If the lumpability hypothesis was not rejected, they compared in the coarser state space (for instance the three-state space {A/G, C, T}), different models corresponding to different high-order dependency hypotheses (more precisely, different mixture transition distribution models, a kind of parametric high-order Markov chains relying on an analogy with autoregressive models). A shortcoming of their approach is that the two stages of analysis are decoupled and that the first stage relies necessarily on a first-order Markov chain assumption. Hence, their approach suffers from a lack of a single statistical model that represents both lumpability and high-order dependencies.

In the remainder of this section, we analyse two different groups of DNA sequences namely (i) intronic sequences and (ii) gene untranslated regions (5' and 3' UTRs). Although both types of sequences are part of genes encoding proteins, they do not by themselves code for protein segments. The introns correspond to the RNA fragments that are excised from the nascent RNA to produce the final messenger RNA. The 3' UTR is the terminal part of the messenger RNA situated downstream of the stop codon (end of the protein coding sequence). Numerous studies have shown that 3' UTR sequences play crucial roles in post-transcriptional regulation of gene expression. These sequences can be recognized by protein complexes that control the stability (Mitchell and Tollervey, 2001), the subcellular localization (Jansen, 2001) or the translation (Macdonald, 2001) of the mRNA molecule. On the opposite, intronic sequences generally do not contain regulatory sequence motifs (except in the vicinity of the 5' and 3' ends).

It should be noted that lumped processes corresponding to different partitions (including at least {A/G, C, T}, {A, C/T, G} and {A/T, C, G}), different lumpability properties and different Markov chain orders were systematically compared while only the most relevant results are reported in the remainder of this section.

6.1 Human introns

The sample comprises 29 intronic sequences of cumulated length 18084. These 29 introns belong to four contiguous human genes (introns 3 to 11 of gene TAP1, introns 1 to 10 of gene TAP2, introns 1 to 5 of gene LMP7, introns 1 to 5 of gene DOB; accession number X66401) located in regions of medium G+C content (45 %).

The estimated transition probability matrix for the original four-state space (Table 2) shows

clearly that the two rows corresponding to the purines A and G are close while the corresponding columns are very different (because of the rare C→G transition with $\widehat{p}_{CG} = 0.05$). This suggests lumping A and G into a single state A/G that denotes purines. The estimated transition probability matrix for the state space {A/G, C, T} is given in Table 3. As expected, for this three-state space, the BIC favors the two-state-dependent lumped process corresponding to the row equality criterion alone while the column ratio criterion eliminates the state-dependent lumped process (Table 4). The rules of thumb of Jeffreys (1961, Appendix B; see also Kass and Raftery (1995)) suggest that a difference of BIC of at least $2 \log 100 = 9.2$ is needed to deem the model with the higher BIC substantially better. The BIC analysis confirms that the output-state-dependent lumped process corresponding to the row sum criterion is overparameterized since for each n (either A or G), $\widehat{b}_{A,A/G}(n) \simeq \widehat{b}_{G,A/G}(n)$; see Table 5.

For second-order Markov chains, the estimated transition probability matrix for the original four-state space is given in Table 6 while the estimated transition probability matrix for the state space {A/G, C, T} is given in Table 7. It should be noted that for the original four-state space, the BIC favors the first-order Markov chain while in the case of the grouping of the purines A and G, the BIC favors the two-state-dependent lumped process based on a second-order Markov chain (Table 4). This means in particular that rows A A, G A, A G, G G | C A, C G | T A, T G | A C, G C | A T, G T respectively tend to be close together (and closer than rows • A | • C | • G | • T respectively); the groups of rows are separated by '|'. The BIC favors the two-state-dependent lumped process irrespective of Markov chain order (Table 4).

If the DNA strand is read in the opposite 3' to 5' direction, the rare C→G transitions are replaced by rare G→C transitions. Hence, G cannot be grouped with other states. In this case, the estimated transition probability matrix for the original four-state space (Table 8) shows clearly that the two rows corresponding to the pyrimidines C and T are close while the corresponding columns are very different. As expected, for this three-state space, the BIC favors the two-state-dependent lumped process irrespective of Markov chain order (Table 9) while the output-state-dependent lumped process is overparameterized; see the observation probabilities in Table 10. It should be noted that for the original four-state space, the BIC favors the first-order Markov chain while in the case of the grouping of the pyrimidines C and T, the BIC favors the two-state-dependent lumped process based on a second-order Markov chain (Table 9).

6.2 3' UTR sequences

The sample comprises 18 human UTR sequences of length ranging from 1.1 to 3.8 kb and of cumulated length 37084, retrieved from UTRdb (Pesole *et al.*, 2002) (accession numbers: 3HSA016017, 3HSA016060, 3HSA016075, 3HSA020009, 3HSA016076, 3HSA012593, 3HSA000005, 3HSA020017, 3HSA016077, 3HSA011748, 3HSA000007, 3HSA011753, 3HSA012606, 3HSA014832, 3HSA012609, 3HSA020024, 3HSA020025, 3HSA016120).

Both in the conventional 5' to 3' direction and in the opposite 3' to 5' direction, the BIC favors the output-state-dependent lumped process based on a second-order Markov chain for the partition {A/T, C, G}; see Tables 11 to 14. The analysis enables to highlight a reversible

grouping of A and T. It should be noted that reversibility can only be stated formally for state-dependent lumped processes based on first-order Markov chains.

7 Concluding remarks

On the basis of the mouse gene introns (of T-cell receptor α/δ-locus and αA-crystallin gene) studied by Raftery and Tavaré (1994), the grouping of the purines A and G can be identified in the conventional 5' to 3' direction and the grouping of the pyrimidines C and T can be identified in the opposite 3' to 5' direction (modeled in all the cases by two-state-dependent lumped processes based on first-order Markov chains corresponding to the row equality criterion alone); the result are not reported.

For the intronic sequences, the grouping of the purines A and G in the conventional 5' to 3' direction entails, on the complementary strand, the grouping of the pyrimidines C and T in the 3' to 5' direction, and conversely, the grouping of the pyrimidines C and T in the opposite 3' to 5' direction entails, on the complementary strand, the grouping of the purines A and G in the 5' to 3' direction. For the 3' UTR sequences, the grouping of A and T in the conventional 5' to 3' direction entails, on the complementary strand, the grouping of A and T in the 3' to 5' direction, and conversely, the grouping of A and T in the opposite 3' to 5' direction entails, on the complementary strand, the grouping of A and T in the 5' to 3' direction. The aggregations found in the conventional 5' to 3' direction and in the opposite 3' to 5' direction are therefore conserved in the complementary strand both for intronic and 3' UTR sequences.

The different possible groupings were evaluated on several human introns. The symmetry property of the two complementary strands is systematically verified on these supplementary examples. The most frequently selected lumped processes are two-state-dependent lumped processes for the grouping of the purines A and G in the conventional 5' to 3' direction (and the grouping of the pyrimidines C and T in the opposite 3' to 5' direction). The different possible groupings were also evaluated on the individual 3' UTR sequences (length between 1 kb and 4 kb) analysed all together in Section 6.2. In addition to the grouping of A and T (in both directions), the grouping of the purines A and G in the conventional 5' to 3' direction and the grouping of the pyrimidines C and T in the opposite 3' to 5' direction are in most cases possible (results not reported). Since the grouping of A and T is always possible for 3' UTR sequences and almost never possible for intronic sequences, this grouping could be tested as a discriminant rule between intronic and 3' UTR sequences. The interpretation of the groupings identified in intronic and 3' UTR sequences in terms of biological processes is an open question. It is known that UTRs contain sequence motifs involved in the regulation of mRNA stability, transport and translation (Mitchell and Tollervey, 2001; Jansen, 2001; Macdonald, 2001). Intronic sequences rarely contain such regulatory sequence motifs. It is thus likely that lumped processes have been able to reveal statistical differences related to these UTR- or intron-specific sequence properties.

The main interest of the proposed approach to lumpability is that different processes corresponding to different lumpability hypotheses and possibly to different high-order dependency hypotheses may be compared by usual measures such as BIC since they are all defined on the

same original state space. Lumped processes and variable-order Markov chains are two complementary ways for defining structurally rich classes of statistical models with reference to (high-order) Markov chains. Hence, lumped processes based on variable-order Markov chains may provide a general framework for the analysis of local dependencies in stationary sequences. Lumped processes based on variable-order Markov chains that provide a more refined modelling of local dependencies compared to simple fixed-order Markov chains may be useful as input of methods for comparing different types of biological sequences (e.g. introns such as discussed in Hall *et al.* (1998)) or as a basis for sophisticated methods for analysing words in sequences such as reviewed in Reinert *et al.* (2000).

The definition of lumped processes from (possibly high-order) Markov chains is similar to the definition of variable-order Markov chains in the sense that new classes of parsimonious models are defined on the basis of lumping mechanisms (of states, of memories). Hence, minimal parameterizations of specific classes of Markov chains are in this way derived and global model selection methods such as BIC apply equally well to these structurally richer families of models as to fixed-order Markov chains. In this more general framework, new difficulties may appear due to the potentially very large number of candidate models to be compared.

Interesting perspectives will be to apply the proposed approach to many other DNA sequences and to test if lumped processes can serve as a predictive tool for the annotation of intronic and UTR sequences.

Acknowledgments

The authors thank Simon Tavaré for kindly providing the mouse intron data, Avner Bar-Hen and Dominique Cellier for their helpful comments, and Didier Piau and a referee for their insightful comments that led to an improvement in the presentation of this paper.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716-723.

Billingsley, P. (1961). Statistical methods in Markov chains. *Annals of Mathematical Statistics*, 32, 12-40.

Bühlmann, P., and Wyner, A. J. (1999). Variable length Markov chains. *The Annals of Statistics*, 27(2), 480-513.

Burke, C. J., and Rosenblatt, M. (1958). A Markovian function of a Markov chain. *Annals of Mathematical Statist*ics, 29, 1112-1122.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*, 2nd edn. New York: Springer.

Csiszár, I., and Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *The Annals of Statistics*, 28(6), 1601-1619.

Ephraim, Y., and Merhav, N. (2002). Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6), 1518-1569.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Volume 1, 3rd edn. New York: Wiley.

Guttorp, P. (1995). *Stochastic Modeling of Scientific Data*. London: Chapman & Hall.

Hall, D. L., Kadafar, K., and Malkinson, A. M. (1998). Statistical methodology for assessing homology of intronic regions of genes. *The Canadian Journal of Statistics*, 26(3), 455-465

Jansen, R. P. (2001). mRNA localization: message on the move. *Nature Reviews Molecular Cell Biology*, 2, 247-256.

Jeffreys, H. (1961). *Theory of Probability*, 3rd edn. Oxford: Oxford University Press.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Katz, R. W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics*, 23(3), 243-249.

Kemeny, J. G., and Snell, J. L. (1976). *Finite Markov Chains*. New York: Springer.

Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems*. London: Chapman & Hall.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Oxford University Press.

Macdonald, P. (2001). Diversity in translational regulation. *Current Opinion in Cell Biology*, 13, 326-331.

Mächler, M. and Bülhmann, P. (2004). Variable length Markov chains: Methodology, computing and software. *Journal of Computational and Graphical Statistics*, 13(2), 435-455.

Mitchell, P., and Tollervey, D. (2001). mRNA turnover. *Current Opinion in Cell Biology*, 13, 320-325.

Pesole, G., Liuni, S., Grillo, G., Licciulli, F., Mignone, F., Gissi, C., and Saccone C. (2002). UTRdb and UTRsite: specialized database of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research*, 30, 335-340.

Prum, B., Rodolphe, F., and de Turckheim, E. (1995). Finding words with unexpected frequencies in DNA sequences. *Journal of the Royal Statistical Society Series* B, 57, 205-220.

Raftery, A. E., and Tavaré, S. (1994). Estimation and modelling repeated patterns in high order Markov chains with the mixture transition distribution model. *Applied Statistics*, 43(1), 179-199.

Reinert, G., Schbath, S., and Waterman, M. S. (2000). Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology*, 7(1/2), 1-46.

Robin, S., and Daudin, J. J. (1999). Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, 36 179-193.

Rogers, D. F., and Plante, R. D. (1993). Estimating equilibrium probabilities for band diagonal Markov chains using aggregation and disaggregation techniques. *Computers & Operations Research*, 20, 857-877.

Ron, D., Singer, Y., and Tishby, N. (1996). The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25, 117-149.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.

Smyth, P., Heckerman, D., and Jordan, M. I. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9, 227-269.

Stefanov, V. T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: An algorithmic approach. *Journal of Applied Probability*, 40, 881-892.

Thomas, M. U., and Barr, D. R. (1977). An approximate test of Markov chain lumpability. *Journal of the American Statistical Association*, 72, 175-179.

Weinberger, M. J., Rissanen, J. J., and Feder, M. (1995). A universal finite memory source. *IEEE Transactions on Information Theory*, 41(3), 643-652.

Table 1. Number of of free parameters of four-state Markov chains as a function of their order.

r	0	1	2	3	4	5	6	7
No. of free param	3	12	48	192	768	3072	12288	49152

Table 2. Human introns (5' to 3'): estimated transition probability matrix for a first-order Markov chain.

	A	C	G	T
A	0.25	0.19	0.28	0.28
C	0.29	0.29	0.05	0.37
G	0.26	0.19	0.29	0.26
T	0.17	0.25	0.25	0.33

Table 3. Human introns (5' to 3'): estimated transition probability matrix for the first-order Markov chain defined on the state space {A/G, C, T}.

	A/G	C	T
A/G	0.54	0.19	0.27
C	0.34	0.29	0.37
T	0.42	0.25	0.33

Table 4. Human introns (5' to 3'): BIC analysis (cumulated sequence length: 18084).

	No. of free param	$2 \log L$	BIC
state space {A, C, G, T}			
first-order Markov chain	12	-48321	-48438.6
second-order Markov chain	48	-47985.9	-48456.4
state space {A/G, C, T} - first-order Markov chain			
state-dependent lumped process	7	-49187.6	-49256.3
two-state-dependent lumped process	9	-48323.9	-48412.1
output-state-dependent lumped process	10	-48323.9	-48421.9
state space {A/G, C, T} - second-order Markov chain			
state-dependent lumped process	19	-48960.2	-49146.4
two-state-dependent lumped process	21	-48096.4	-48302.2
output-state-dependent lumped process	22	-48096.4	-48312

Table 5. Human introns (5' to 3'): estimated observation probabilities.

	A	G
two-state-dependent lumped process		
A/G → A/G	0.48	0.52
C → A/G	0.86	0.14
T → A/G	0.41	0.59
output-state-dependent lumped process		
A → A/G	0.48	0.52
C → A/G	0.86	0.14
G → A/G	0.48	0.52
T → A/G	0.41	0.59

Table 6. Human introns (5' to 3'): estimated transition probability matrix for a second-order Markov chain.

	A	C	G	T
A A	0.32	0.17	0.25	0.26
C A	0.19	0.21	0.28	0.32
G A	0.28	0.17	0.34	0.21
T A	0.23	0.22	0.23	0.32
A C	0.37	0.26	0.05	0.32
C C	0.27	0.29	0.05	0.39
G C	0.3	0.29	0.08	0.33
T C	0.25	0.31	0.03	0.41
A G	0.31	0.17	0.29	0.23
C G	0.21	0.26	0.28	0.25
G G	0.27	0.2	0.29	0.24
T G	0.23	0.19	0.28	0.3
A T	0.2	0.22	0.26	0.32
C T	0.14	0.29	0.26	0.31
G T	0.2	0.22	0.32	0.26
T T	0.17	0.26	0.19	0.38

Table 7. Human introns (5' to 3'): estimated transition probability matrix for a second-order Markov chain defined on the state space {A/G, C, T}.

		A/G	C	T
A/G	A/G	0.58	0.18	0.24
C	A/G	0.48	0.21	0.31
T	A/G	0.49	0.2	0.31
A/G	C	0.4	0.28	0.32
C	C	0.32	0.29	0.39
T	C	0.28	0.31	0.41
A/G	T	0.49	0.22	0.29
C	T	0.4	0.29	0.31
T	T	0.36	0.26	0.38

Table 8. Human introns (3' to 5'): estimated transition probability matrix for a first-order Markov chain.

	A	C	G	T
A	0.26	0.28	0.24	0.22
C	0.19	0.29	0.18	0.34
G	0.31	0.05	0.28	0.36
T	0.21	0.28	0.18	0.33

Table 9. Human introns (3' to 5'): BIC analysis (cumulated sequence length: 18084).

	No. of free param	$2 \log L$	BIC
state space {A, C, G, T}			
first-order Markov chain	12	-48321	-48438.6
second-order Markov chain	48	-47978.6	-48449.2
state space {A, C/T, G} - first-order Markov chain			
state-dependent lumped process	7	-49189.6	-49258.2
two-state-dependent lumped process	9	-48328.4	-48416.6
output-state-dependent lumped process	10	-48328.4	-48426.4
state space {A, C/T, G} - second-order Markov chain			
state-dependent lumped process	19	-48930.8	-49117.1
two-state-dependent lumped process	21	-48069.7	-48275.5
output-state-dependent lumped process	22	-48069.7	-48285.3

Table 10. Human introns (3' to 5'): estimated observation probabilities.

	C	T
two-state-dependent lumped process		
A → C/T	0.56	0.44
C/T → C/T	0.46	0.54
G → C/T	0.13	0.87
output-state-dependent lumped process		
A → C/T	0.56	0.44
C → C/T	0.46	0.54
G → C/T	0.13	0.87
T → C/T	0.46	0.54

Table 11. 3' UTR sequences (5' to 3'): BIC analysis (cumulated sequence length: 37084).

	No. of free param	$2 \log L$	BIC
state space {A, C, G, T}			
first-order Markov chain	12	-99844.5	-99970.8
second-order Markov chain	48	-99103.6	-99608.6
state space {A/T, C, G} - first-order Markov chain			
output-state-dependent lumped process	10	-99860.8	-99966
state space {A/T, C, G} - second-order Markov chain			
output-state-dependent lumped process	22	-99347.1	-99578.6

Table 12. 3' UTR sequences (5' to 3'): estimated observation probabilities for an output-state-dependent lumped process.

	A	T
A $\rightarrow$ A/T	0.54	0.46
C $\rightarrow$ A/T	0.47	0.53
G $\rightarrow$ A/T	0.5	0.5
T $\rightarrow$ A/T	0.38	0.62

Table 13. 3' UTR sequences (3' to 5'): BIC analysis (cumulated sequence length: 37084).

	No. of free param	$2 \log L$	BIC
state space {A, C, G, T}			
first-order Markov chain	12	-99843.9	-99970.2
second-order Markov chain	48	-99101.5	-99606.5
state space {A/T, C, G} - first-order Markov chain			
output-state-dependent lumped process	10	-99873.8	-99979
state space {A/T, C, G} - second-order Markov chain			
output-state-dependent lumped process	22	-99359.3	-99590.8

Table 14. 3' UTR sequences (3' to 5'): estimated observation probabilities for an output-state-dependent lumped process.

	A	T
A $\rightarrow$ A/T	0.57	0.43
C $\rightarrow$ A/T	0.45	0.55
G $\rightarrow$ A/T	0.46	0.54
T $\rightarrow$ A/T	0.41	0.59

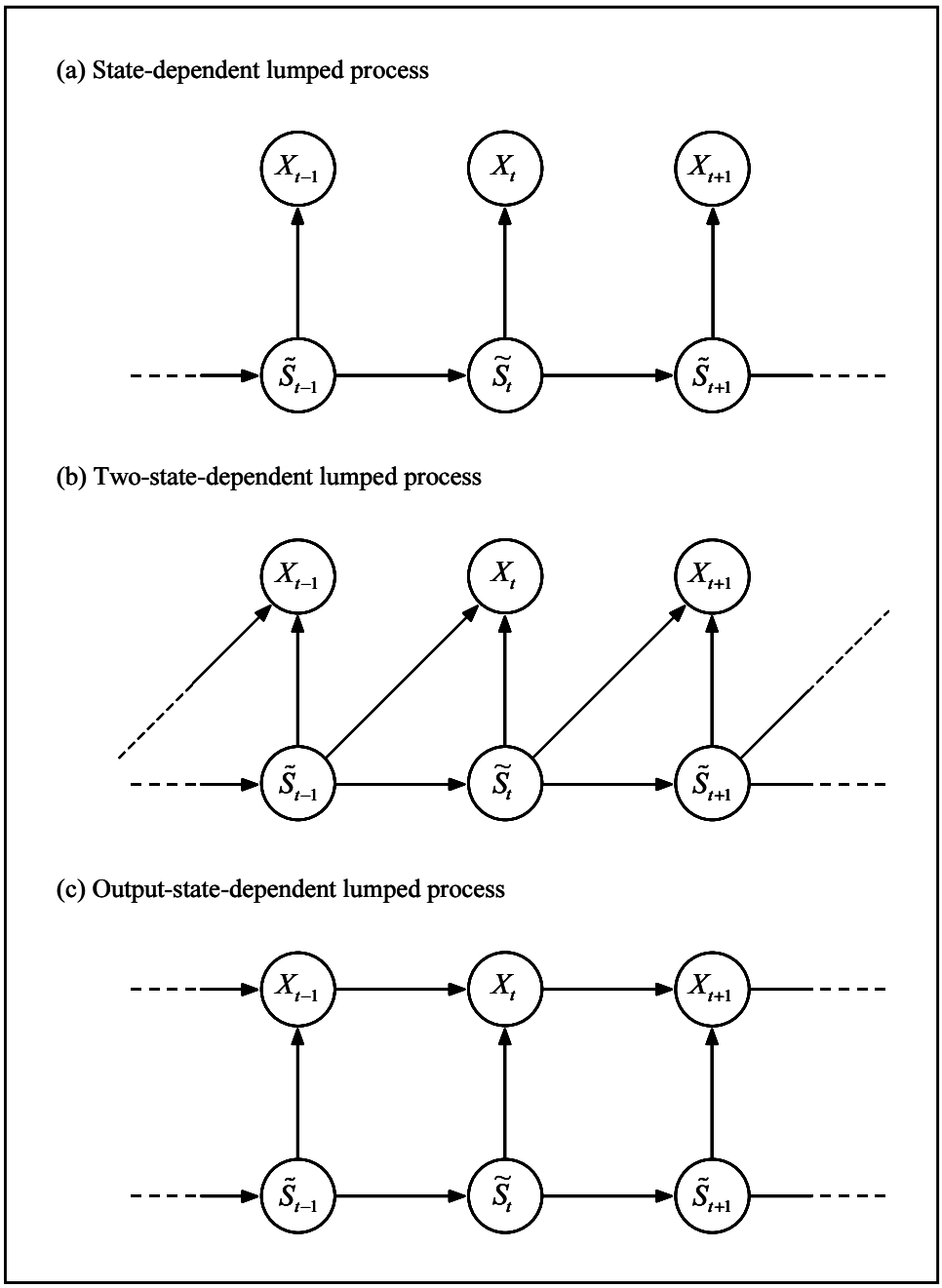

Figure 1. Conditional independence graphs of lumped processes constructed from first-order Markov chains.

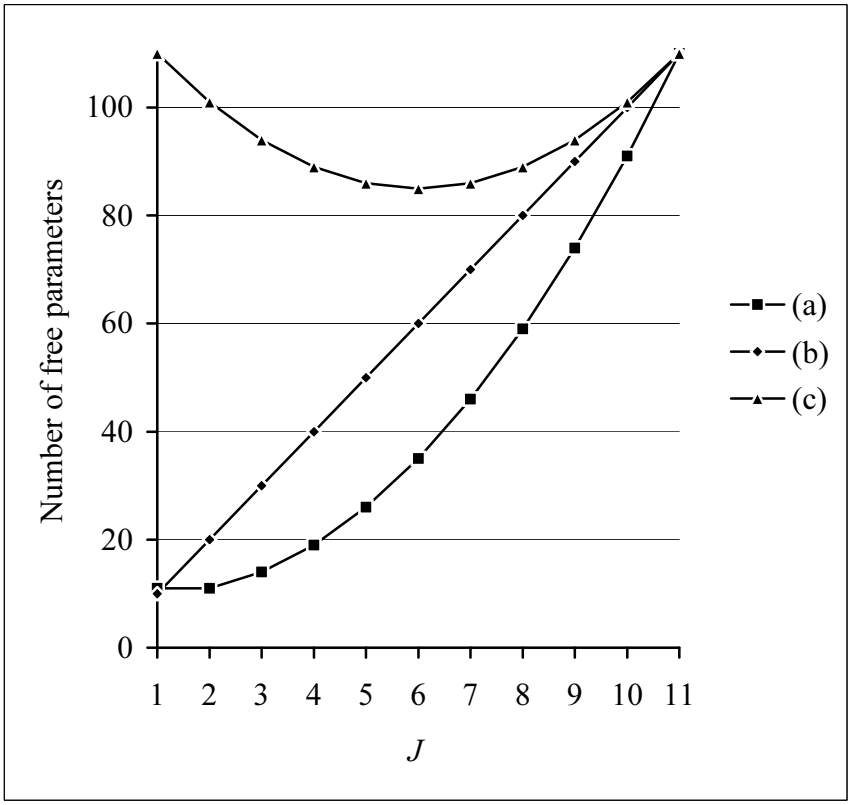

Figure 2. Number of free parameters as a function of the number of states J for $N = 11$: (a) $\varphi_N(J)$ (state-dependent lumped process), (b) $\phi_N(J)$ (two-state-dependent lumped process), (c) $\psi_N(J)$ (output-state-dependent lumped process).

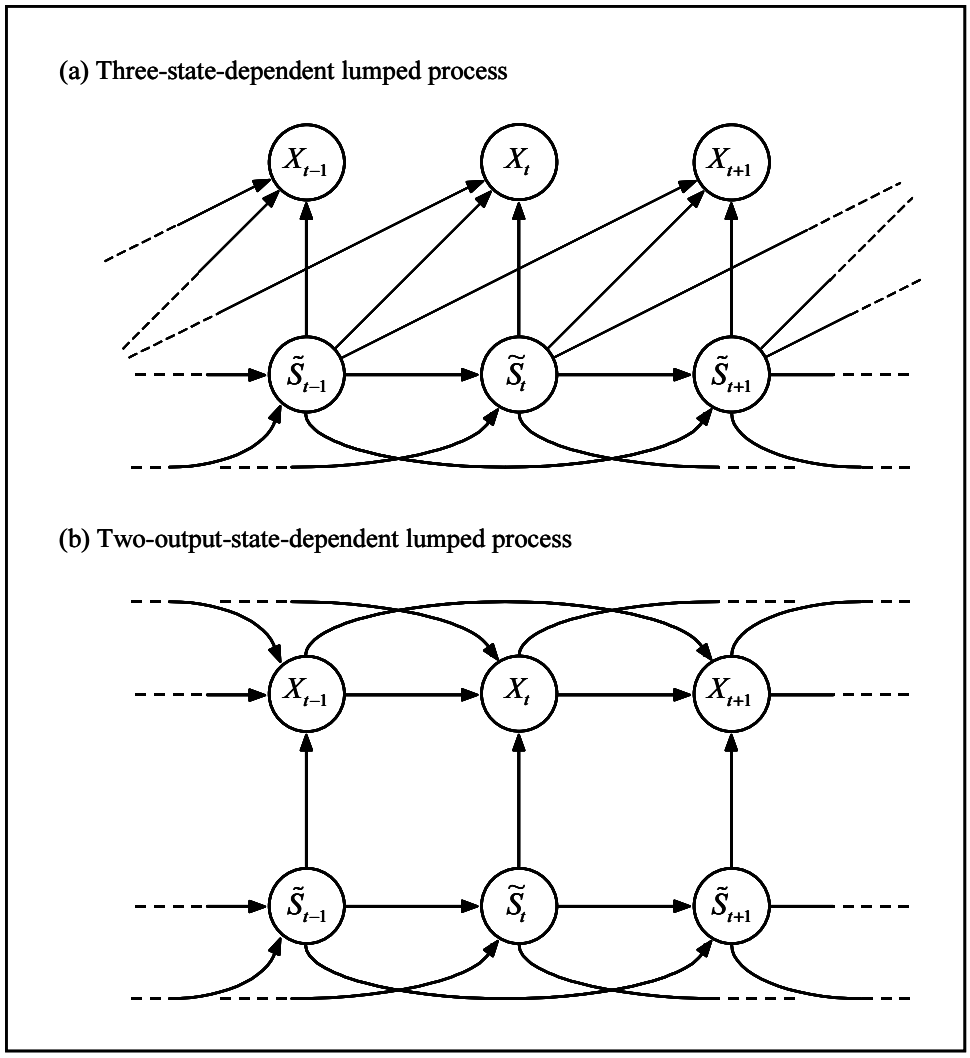

Figure 3. Conditional independence graphs of lumped processes constructed from second-order Markov chains.